Using Biased Discriminant Analysis for Email Filtering

Juan Carlos Gomez[1] and Marie-Francine Moens[2]

[1] ITESM, Eugenio Garza Sada 2501, Monterrey NL 64849, Mexico
juancarlos.gomez@invitados.itesm.mx
[2] Katholieke Universiteit Leuven, Celestijnenlaan 200A, B-3001 Heverlee, Belgium
sien.moens@cs.kuleuven.be

Abstract. This paper reports on email filtering based on content features. We test the validity of a novel statistical feature extraction method, which relies on dimensionality reduction to retain the most informative and discriminative features from messages. The approach, named Biased Discriminant Analysis (BDA), aims at finding a feature space transformation that closely clusters positive examples while pushing away the negative ones. This method is an extension of Linear Discriminant Analysis (LDA), but introduces a different transformation to improve the separation between classes and it has up till now not been applied for text mining tasks.
We successfully test BDA under two schemas. The first one is a traditional classification scenario using a 10-fold cross validation for four ground truth standard corpora: LingSpam, SpamAssassin, Phishing corpus and a subset of the TREC 2007 spam corpus. In the second schema we test the anticipatory properties of the statistical features with the TREC 2007 spam corpus.
The contributions of this work is the evidence that BDA offers better discriminative features for email filtering, gives stable classification results notwithstanding the amount of features chosen, and robustly retains their discriminative value over time.

1 Introduction.

In data mining the goal is to find previously unknown patterns and relations in large databases [12], and a common task is automatic classification. Here, given a set of instances with known values of their attributes and their classes, the aim of this task is to predict automatically the class of a new instance, when only the values of its features are known.

When classifying email messages, commonly the data contained in them are very complex, multidimensional or represented by a large number of features. Since when using many features, we need a corresponding increase in the number of annotated examples to train from to ensure a correct mapping between the features and the classes [1][5], the use of any kind of dimensionality reduction method is useful in the classification task.

In this paper, we want to discriminate between two classes of email messages (spam or phishing from ham) with the purpose of detecting potentially dangerous emails. In order to do it, we advocate an approach which is an extension of the traditional *Linear Discriminant Analysis* (LDA), named *Biased Discriminant Analysis* (BDA) [3][15]. This method is especially suited to find a feature space transformation that closely clusters the positive examples while pushing away the negative ones. The method has never been used in text mining tasks, but seems promising in binary classification such as spam and phishing mail filtering.

We test this algorithm under two scenarios. First, we try to obtain the smallest number of features to have a good performance in classification. In that case, these features can be understood as the core profile of a data set, which allows a fast training of classifiers. Second, we test the capability of these core profiles of persisting over time, in order to anticipate new dangerous messages, we do this by ordering emails by date and by training our method on older emails and testing on more recent ones. Both schemata are evaluated based on standard datasets for spam and phishing mail filtering.

We compare the BDA results with the basic LDA method to see the improvements included with the more recent algorithm. Additionally, for comparison we present the results for a classifier trained with the complete vocabulary composed by all the unique terms in each data set.

The contribution of our work is the evidence that the technique of Biased Discriminant Analysis offers excellent discriminative features for the filtering, that gives stable results notwithstanding the amount of features chosen, and that robustly retains their discriminative value over time. Our findings contribute to the development of more advanced email filters and open new opportunities for text classification in general.

The remainder of this paper is organised as follows. Section 2 overviews related work on dimensionality reduction. Section 3 introduces the model for dimensionality reduction using BDA. Section 4 discusses our experimental evaluation of the method for email classification. Section 5 concludes this work with lessons learnt and future research directions.

2 Related Work.

Several methods have been proposed for email filtering [13], and one of the most promising approaches is the use of content-based filters [19], which use the text content of the messages rather than black lists, header or sender information. In this sense, machine learning and data mining methods are especially attractive for this task, since they are capable of adapting to the evolving features of spam and phishing messages' content. There is plenty of work devoted to email filtering [18], including some seminal papers for spam classification using traditional Bayesian filters like [2]. There are also interesting works on phishing detection like [11], describing a set of features to distinguish phishing emails. Nevertheless most of the methods devoted to email filtering use bag-of-words as features and Bayesian methods to perform the classification. Recently, works like [7] and

[16] where the authors use compression models and n-grams to produce more robust features and more sophisticated classifiers like Support Vector Machines, are starting to emerge. Our work pretends to contribute with a new approach in the task of email classification specifically focusing on feature extraction.

Dimensionality reduction has been popular since the early 90s in text processing tasks, like, for example, the technique of Latent Semantic Analysis (LSA) [10]. LSA is an application of principal component analysis where a document is represented along its semantic axes or topics. In a text categorization task, documents are represented by a LSA vector model both when training and testing the categorization system. However, these models do not exploit class information.

Probabilistic topic models such as probabilistic Latent Semantic Analysis (pLSA) [14] and Latent Dirichlet Allocation (LDrA) [6] are currently popular as topic representation models. Documents are represented as a mixture of topic distributions and topics as a mixture of word distributions. However, in text categorization, information about the text categories is not taken into account. Very recently, the LDrA model has been used for spam classification [4].

Linear Discriminant Analysis (LDA) uses class information in order to separate well the classes. Recently, the computer vision community has successfully proposed several variants of LDA that artificially pull apart the positive and the negative examples. One of these is *Biased Discriminant Analysis* (BDA), used by Huang et al. [15] for learning a better ranking function based on relevance feedback in image search. To our knowledge, this LDA variant is only recently developed and has not been used for text classification or email filtering.

BDA is an eigenvalue based method. An eigenvalue is a number indicating the weight of a particular pattern or cluster expressed by the corresponding eigenvector. The larger the eigenvalue the more important the pattern is.

3 Biased Discriminant Analysis.

The final goal of this work is to classify email messages in a priori defined mutual exclusive classes. Our concrete aim is to transform the original feature space of emails into a less dimensional space. This space would be expressed in terms of statistical similarities and differences between messages. The new space is intended to be easier to deal with because of its size, carrying the most important part (core profiles) of the information needed to filter emails.

Let $\{(\mathbf{x}_1, c_1), (\mathbf{x}_2, c_2), \ldots, (\mathbf{x}_n, c_n)\}$ be a set of email messages with their corresponding classes, where $\mathbf{x}_i \in \mathbb{R}^d$ is the i-th email, represented by a d dimensional row vector, and $c_i \in \mathbf{C}$ is the class of $\mathbf{x}_i$. In this work we have $\mathbf{C} = \{-1, +1\}$, where -1 refers to the negative class N (ham messages) and $+1$ to the positive class P (spam or phishing).

The goal of the data dimensionality reduction is to learn a $d \times l$ projection matrix $\mathbf{W}$, which can project to:

$$\mathbf{z}_i = \mathbf{x}_i \mathbf{W} \tag{1}$$

where $\mathbf{z}_i \in \mathbb{R}^l$ is the projected data with $l \ll d$, such that in the projected space the data from different classes can be effectively discriminated.

As was mentioned before, BDA [15] is a variant of LDA, where BDA seeks to transform the feature space so that the positive examples cluster together and each negative instance is pushed away as far as possible from this positive cluster, resulting in the centroids of both the negative and positive examples being moved. BDA aims at maximizing the following function:

$$\mathbf{W}^* = \arg\max_{\mathbf{W}} \frac{|\mathbf{W}^T \mathbf{S}_{PN} \mathbf{W}|}{|\mathbf{W}^T \mathbf{S}_P \mathbf{W}|} \tag{2}$$

The inter-class (positive-negative) scatter matrix $\mathbf{S}_{PN}$, is computed as follows, where μ_P is the mean of the examples in the positive class:

$$\mathbf{S}_{PN} = \Sigma_{\mathbf{y} \in N} (\mathbf{y} - \mu_P)^T (\mathbf{y} - \mu_P) \tag{3}$$

and the intra-class matrix (positive) scatter matrix $\mathbf{S}_P$, is computed as follows:

$$\mathbf{S}_P = \Sigma_{\mathbf{x} \in P} (\mathbf{x} - \mu_P)^T (\mathbf{x} - \mu_P) \tag{4}$$

We then perform an eigenvalue decomposition on $\mathbf{S}_P^{-1} - \mathbf{S}_{PN}$, and construct the $d \times l$ matrix $\mathbf{W}$ whose columns are composed by the eigenvectors of $\mathbf{S}_P^{-1} - \mathbf{S}_{PN}$ corresponding to its largest eigenvalues.

The goal of BDA is to transform the data set $\mathbf{X}$ into a new data set $\mathbf{Z}$ using the projection matrix $\mathbf{W}$, with $\mathbf{Z} = \mathbf{X}\mathbf{W}$ in such a way the examples inside the new data set are well separated by class. In this case, $\mathbf{Z}$ represents the training data. The test examples, for which we do not know the class, are projected using the matrix $\mathbf{W}$ to be represented in the new BDA space. Then, if $\mathbf{q}$ is a test example, its projection using BDA is $\mathbf{u} = \mathbf{q}\mathbf{W}$.

4 Experimental Results.

The four public email corpora we use for performing our tests are: Ling-Spam (LS) [3] [2], SpamAssassin (SA) [4] , TREC 2007 spam corpus ($TREC$) [5][9] and a subset of Phishing Corpus (PC), created by randomly selecting 1,250 phishing messages from the Nazario's corpus [6] and 1,250 ham messages from the TREC corpus. The number of emails in each corpus is listed in table 1.

For this work, before performing BDA, the emails have to be transformed into vectors. First we remove the structure information, i.e. the header and the HTML tags, to retain only the text content, as seen in figure 1. We focus in this article on content based email filtering and realize that important discriminative information in headers of the email is ignored (e.g., address of the sender). Second, we build a vocabulary by removing stop words and words that are evenly distributed over the classes, the latter using a mutual information statistic, obtaining 5000 initial features. Finally, we weight the remaining words in each document by a TF-IDF schema while representing each message as a vector.

[3] Available at: http://nlp.cs.aueb.gr/software.html
[4] Available at: http://spamassassin.apache.org/publiccorpus/
[5] Available at: http://plg.uwaterloo.ca/ gvcormac/treccorpus07/
[6] Available at: http://monkey.org/ jose/wiki/doku.php?id=PhishingCorpus

```
It is going to be HUGE. Target sym: CDYV, Price (current): $0.089, 5 Day
Target price: $0.425, Action: Strong Buy/Hold.. Here comes the REAL BIG
ONE!.. See the news, catchall, call your broker!!
```

Fig. 1. Typical text content of a spam email from TREC spam corpus. We use only the words inside the messages to perform the BDA.

Corpus	Spam	Phishing	Ham	Total
LS	481		2,412	2,893
SA	1,897		4,150	6,047
TREC	50,199		25,220	75,419
PC		1,250	1,250	2,500
Total	52,576	1,250	33,032	

Table 1. Number of messages per corpus.

The training model is constructed by applying the BDA to the message vectors of the training set. Then, a classifier is trained based on this data and tested using the new messages expressed also as vectors in the BDA space. Experimentally we decided to use the bagging ensemble classifier [8], using as single classifier the C4.5 decision tree [17]. The rationale for this is that C4.5 is fast and easy to train and has good performance with small number of features, and that bagging presents a general well behavior by weighting the results of the trees and by reducing the variance of the data set and the overfitting.

Because the feature extraction results in a ranked list of features, we perform experiments by considering the 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, and 2048 highest ranked features for each method BDA and LDA and then we compare all the results. Additionally, we compare with a baseline method (AUT) which uses all unique terms from each training set in the classification.

We present results using the area under the ROC, which aims at a high true positive rate and a low false positive rate. The ROC metric is very important for commercial settings, where the cost for misclassifying a legitimate email as illegal (false positive) is really high. In addition we provide results in terms of overall accuracy of the classification for better understanding the behavior of the algorithms.

Figures 2 and 3 show the results for area under ROC and accuracy for the application of BDA, LDA and the baseline AUT to the four public corpora, performing a normal classification using a 10-fold cross validation for spam and phishing filtering. From these results we can observe that the statistical features extracted by BDA are well suited to discriminate the spam or phishing messages from the ham messages in these corpora. BDA is able to reach good performance almost independently of the number of features used. There are several published works where some of these corpora are used, like [2], [7] and [19]. Nevertheless in these works the authors do not present the complete information about the performance of their methods, and a direct comparison with our results is not possible.

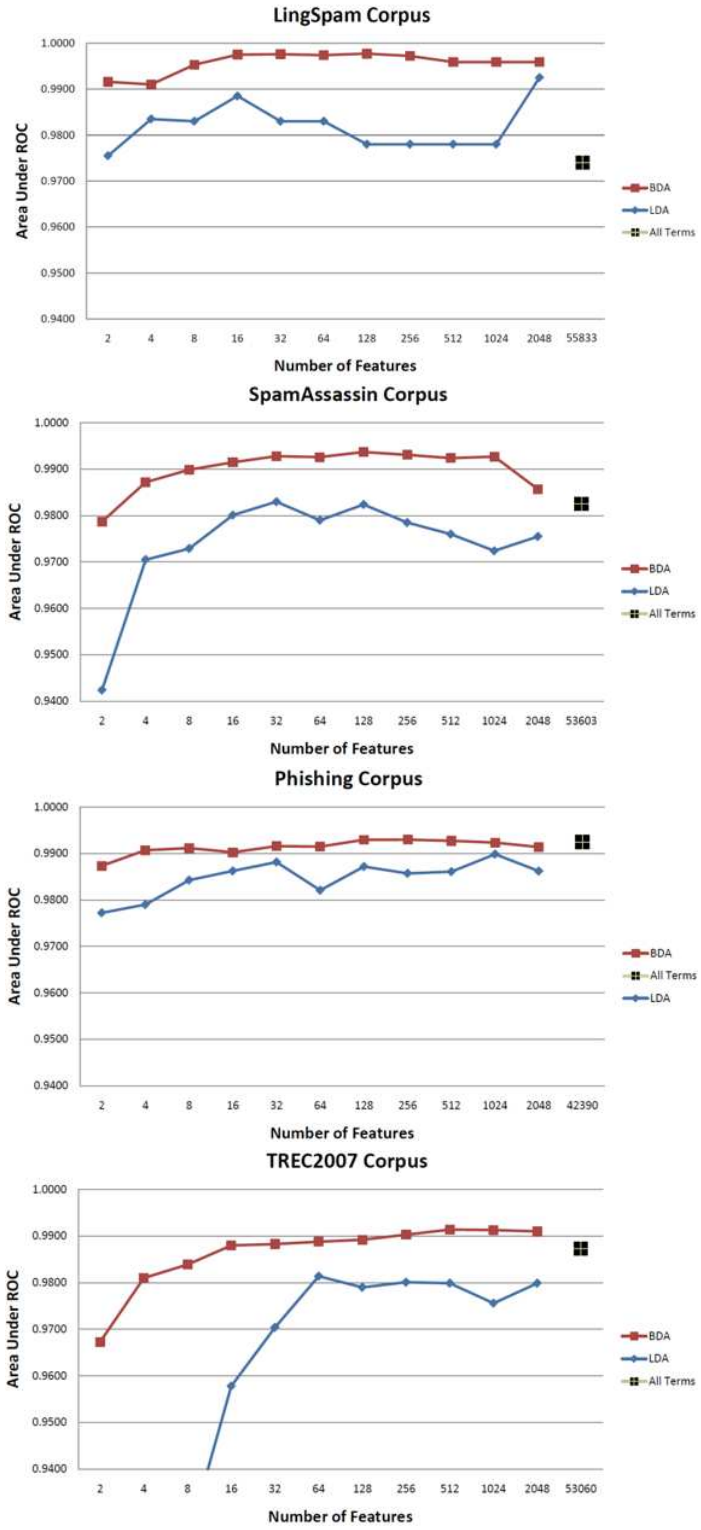

Fig. 2. Performance of algorithms for area under ROC for LS, SA, PC and TREC corpora using 10-fold cross validation.

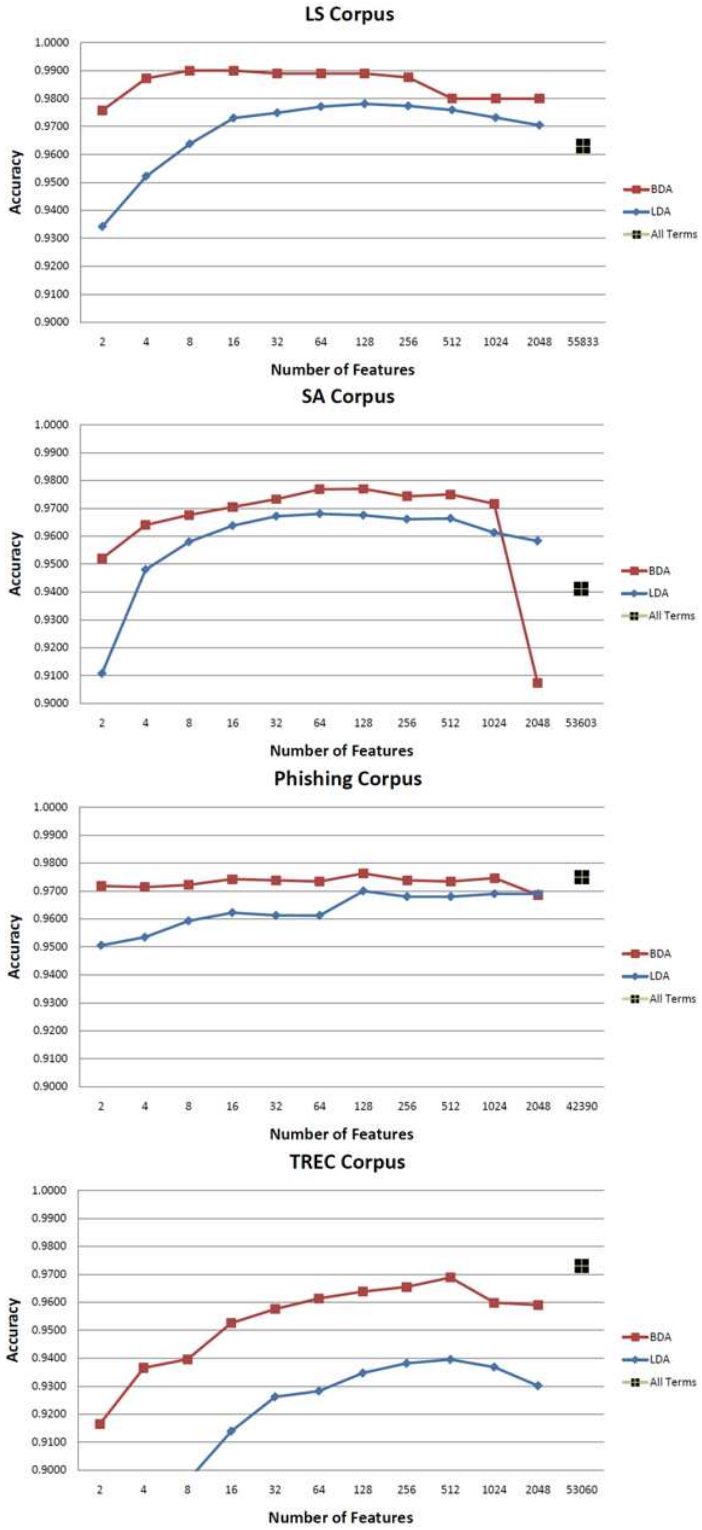

Fig. 3. Performance of algorithms for accuracy for LS, SA, PC and TREC corpora using 10-fold cross validation.

Using 10-fold cross validation BDA performs better than LDA and the baseline AUT in LS and SA corpora for both, area under ROC and accuracy. In the case of PC and TREC corpora, AUT reaches good results, but the number of features used makes it impractical to implement a filter like that. On the other hand, with a much smaller number of features BDA is able to reach competitive (or even better) results with the corresponding minor cost of training and especially of testing a filter.

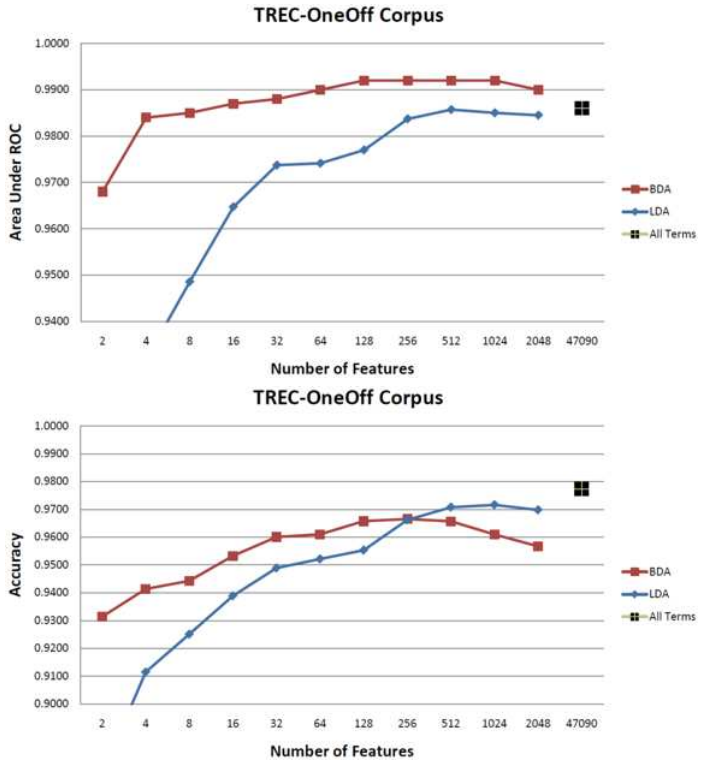

Fig. 4. Performance of algorithms for area under ROC (above) and accuracy (below) for TREC corpus, performing a one-off experiment.

In figure 4 we see the performance of the methods for the one-off experiment for TREC corpus. In this case we want to test the persistency and robustness of the features extracted; then we select a small subset of 9,020 messages, corresponding to the first week of the data set, for training, and the rest 66,399 messages, corresponding to (almost) 11 weeks in the future, for testing. Similar to the previous experiment, the statistical features extracted by BDA reach a good performance with only a small number of features. This means BDA performs quite well when creating robust predictive profiles from past training

data. It is possible to observe that the number of active features is generally low, even if the corpus presents a big variation of topics. The testing of these profiles with completely new data containing unseen messages prove that the statistical features from BDA are able to generalize over a bigger variation inside a data set.

The main aims of this work are the filtering of spam and phishing messages, while performing this filtering with just a few features representing a small set of core profiles and proving the persistency of the core features over time (robustness). As is confirmed by the results presented in this section, we have accomplished these goals.

5 Conclusions.

In this paper we performed content-based email filtering using a statistical feature extraction method. This method, named Biased Discriminant Analysis (BDA), is an approach understood as a dimensionality reduction technique, which especially aims at better discriminating positive from negative examples. The obtained essential statistical features carry very useful information (core profiles of the data set), which is highly discriminative for email classification and robust to persist over time.

The results show a very good classification performance when using BDA for filtering emails in standard benchmarking data sets using 10-fold cross validation and in an anticipatory scenario, where the task was to separate spam and phishing from ham messages. In this sense, BDA is effective for classifying emails and robust when predicting the type of email when trained on older data. Overall, BDA performs excellently and better than standard LDA and (most of the times) than the baseline bag-of-words method, measured in terms of the area under the ROC and accuracy, even when emails are described with very few features. The results obtained with BDA are not very dependent on the number of features chosen, which is an advantage in a text classification task.

Spam filters in practical settings often rely on a small set of signature rules that form a profile. The proposed technique perfectly fits this scenario, by yielding excellent classification results based on a limited number of features. The focus of our work was on content-based filtering. It would be interesting to investigate how our method can be integrated with non-content based features for email filtering such as the provenance of the emails. In the future we want to apply our method to other text classification tasks, possibly customized with multilinear (tensor) methods for dimensionality reduction, that have the possibility to include, for instance, content and non-content features in email filtering.

Acknowledgment

We thank the EU FP6-027600 Antiphish consortium (http://www.antiphishresearch.org/) and in particular Christina Lioma, Gerhard Paaß, André Bergholz, Patrick Horkan, Brian Witten, Marc Dacier and Domenico Dato.

References

1. D. W. Aha, D. F. Kibler, and M. K. Albert. Instance-based learning algorithms. *Machine Learning*, 6:37–66, 1991.
2. I. Androutsopoulos, J. Koutsias, K. V. Chandrinos, K. V. Ch, G. Paliouras, and C. D. Spyropoulos. An evaluation of naive bayesian anti-spam filtering. In *ECML '00: Proceedings of the 11th European Conference on Machine Learning*, pages 9–17, 2000.
3. G. Baudat and F. Anouar. Generalized discriminant analysis using a kernel approach. *Neural Compututation*, 12(10):2385–2404, 2000.
4. B. István, S. Jácint, and B. A. A. Latent dirichlet allocation in web spam filtering. In *AIRWeb '08: Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web*, pages 29–32, 2008.
5. C. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press, 1995.
6. D. M. Blei, A. Y. Ng, M. I. Jordan, and J. Lafferty. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:2003, 2003.
7. A. Bratko, G. Cormack, B. Filipic, T. Lynam, and B. Zupan. Spam filtering using statistical data compression models. *Journal of Machine Learning Research*, 7:2673–2698, 2006.
8. L. Breiman. Bagging predictors. In *Machine Learning*, pages 123–140, 1996.
9. G. V. Cormack. Spam track overview. In *TREC-2007: Sixteenth Text REtrieval Conference*, 2007.
10. S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407, 1990.
11. I. Fette, N. Sadeh, and A. Tomasic. Learning to detect phishing emails. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 649–656, New York, NY, USA, 2007. ACM.
12. K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, London, 1990.
13. J. Goodman, D. Heckerman, and R. Rounthwaite. Stopping spam. *Scientific American*, 292(4):42–88, 2005.
14. T. Hofmann. Probabilistic latent semantic indexing. In *Uncertainty in Artificial Intelligence*, pages 50–57, 1999.
15. T. S. Huang, C. K. Dagli, S. Rajaram, E. Y. Chang, M. I. Mandel, G. E. Poliner, and D. P. W. Ellis. Active learning for interactive multimedia retrieval. *Proceedings of the IEEE*, 96(4):648–667, 2008.
16. I. Kanaris, K. Kanaris, I. Houvardas, and E. Stamatatos. Words vs. character n-grams for anti-spam filtering. *Int. Journal on Artificial Intelligence Tools*, 16(6):1047–1067, 2007.
17. J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, California, 1993.
18. T. S. Guzella and W. M. Caminhas. A review of machine learning approaches to spam filtering. *Expert Systems with Applications*, 36:10206-10222, 2009.
19. B. Yu and Z.-b. Xu. A comparative study for content-based dynamic spam classification using four machine learning algorithms. *Knowledge-Based Systems*, 21(4):355–362, 2008.